AN ANALYSIS OF WEIGHT GAIN AMONG CALL CENTER WORKERS

Elizabeth Anderson, Gokcen Buyukbas, Xiao Li

ABSTRACT

Obesity related diseases are costly for businesses, incentivising interventions to reduce obesity. We examined factors related to weight gain in one specific case of employees working at one call center. First, we examined multiple measures of weight gain, in addition to explanatory variables, such as the time of shift and exercise levels. We used random forest regression and full information maximum likelihood estimation(FIML) to identify associations between predictors. Both of these models agreed that the most important predictors were BMI and age, while the others didn't have a significant effect to predict the pounds gained.

INTRODUCTION

To begin with analysing the data, we started with looking at the descriptive statistics for each variable that can be found in Table 1.

Table 1: Summary statistics

Measurement	Statistic	Measurement	Statistic
Age -Numerical (N=322)	Mean = 33.76 Sd = 9.89 Median = 31	Vigorous exercise time (min/week) -Numerical (N=352)	Mean = 74 Sd = 114 Median = 27
Gender (N=347) -Categorical Female Male	 248 (71%) 99 (29%)	Total exercise time (min/week) -Numerical (N=351)	Mean = 1306 Sd = 1553 Median = 822
BMI -Numerical (N=253)	Mean = 27.82 Sd = 6.1 Median = 27	Weight gain -Categorical (N=348)	No = 111 (32%) Yes = 237 (68%)
Walkin exercise time (min/week) -Numerical (N=352)	Mean = 123 Sd = 212 Median = 60	Pounds gained -Numerical (N=342)	Mean = 11 Sd = 13 Median = 8
Moderate exercise time (min/week) -Numerical (N=351)	Mean = 74 Sd = 140 Median = 30	Shift (N=348) -Categorical 8am 9 am 10 am 11 am 12 pm 1pm 2 pm Other	Count 115 (32.7%) 56 (15.9%) 50 (14.2%) 44 (12.5%) 14 (4%) 8 (2.3%) 15 (4.3%) 15 (4.3%)

There are more female employees in the data (79%), but the distribution of pounds gained are very similar for men and women as seen in Figure 1. The number of employees who gained weight decreases as the amount of gained weight increases, i.e. the distribution of pounds gained

is right-skewed.

Figure 1: Histogram of pounds gained for male and female.

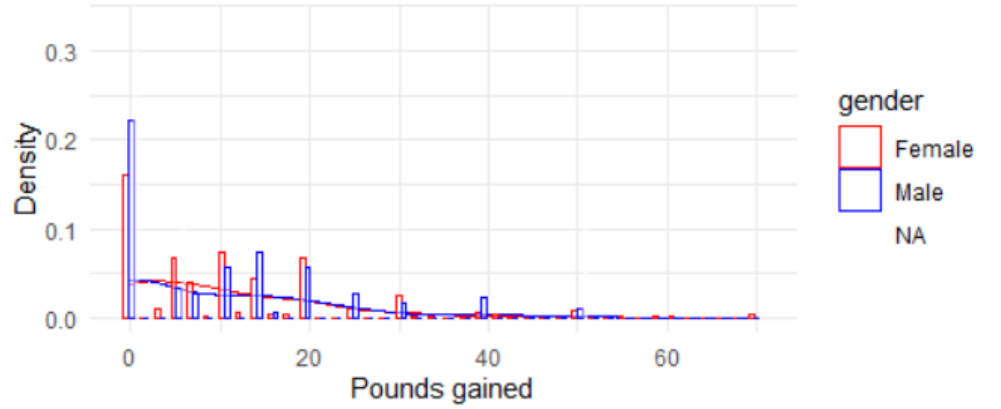


Figure 2 shows the boxplots and the density curves indicating that the people in early morning shifts have gained weight more than later shifts. However, the difference is not highly significant.

Figure 2. Pounds gained vs Shift

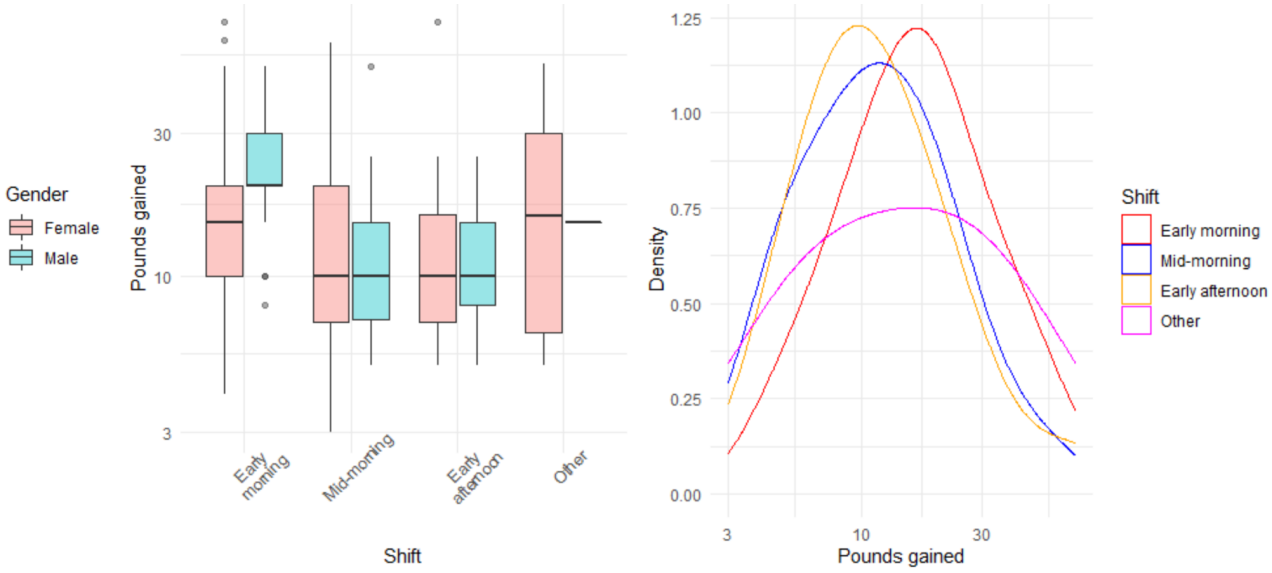
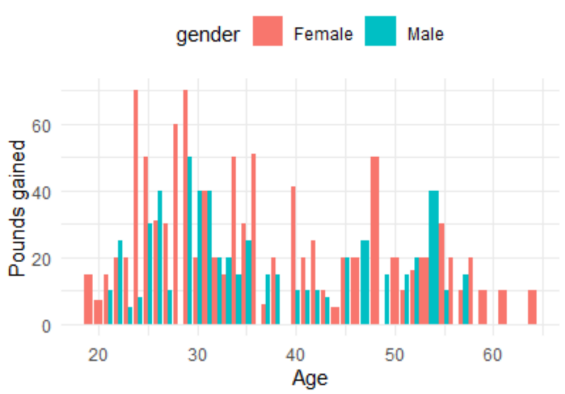


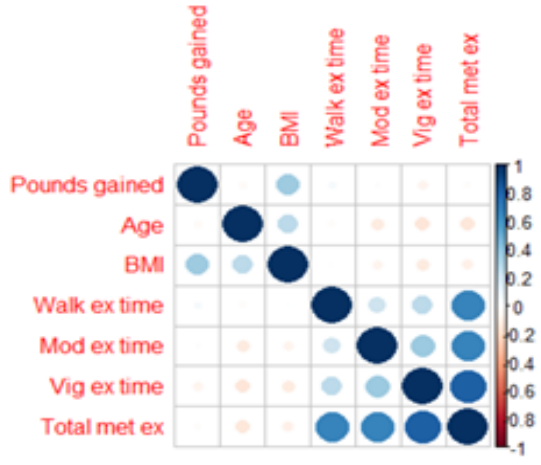
Figure 3: Age vs Pounds gained



When we look at the distribution of age and its relation with the pounds gained (Figure 3), we see that the younger people gained more weight than the older people. Younger women seem to have gained more weight than men, while for 50+, there is not much difference between the genders.

The correlation between the continuous variables can be seen in Figure 4.

Figure 4: Correlation matrix



Exercise times among different types of exercise are highly correlated, as can be expected. However, they don't correlate with the amount of weight gained. On the other hand, BMI and age have a positive correlation with the pounds gained.

Now, let's have a look at the relationship of pounds gained with BMI and the activity time.

We used a logarithmic scale for each variable since they were right-skewed. In Figure 5, we see a null-plot for each exercise category, indicating that there is no significant relationship between the exercise time and pounds gained.

Figure 5: Exercise time vs pounds gained, all variables are in logarithmic scale

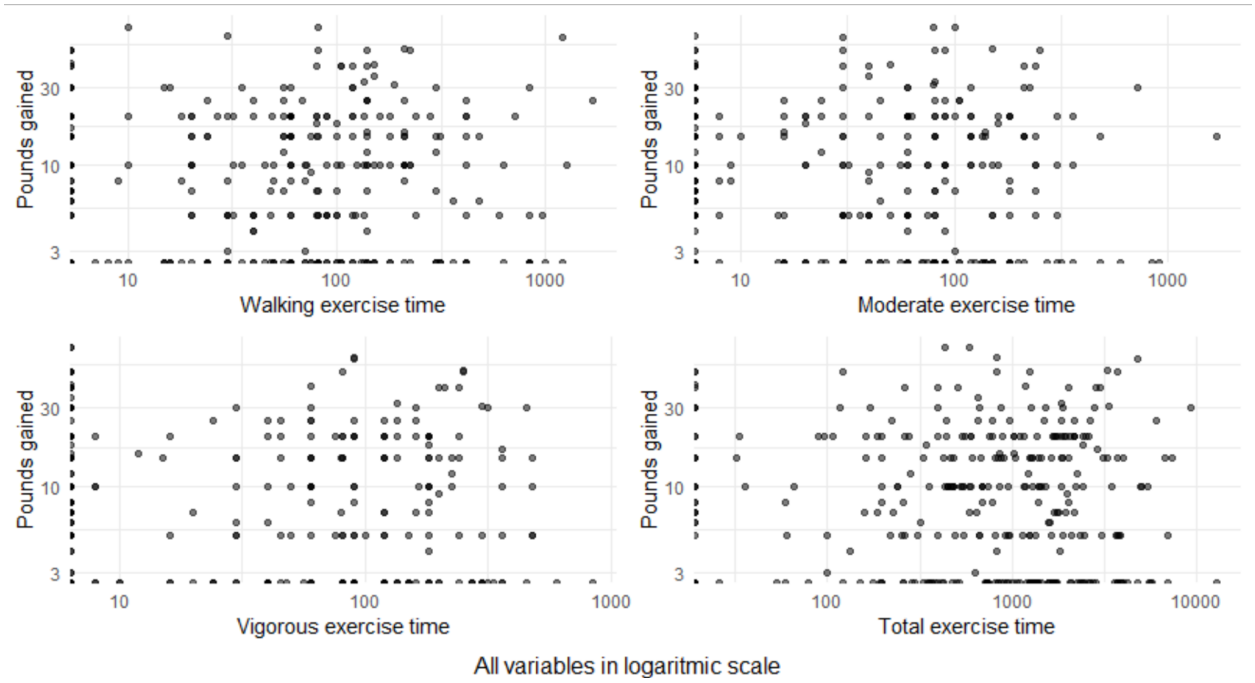
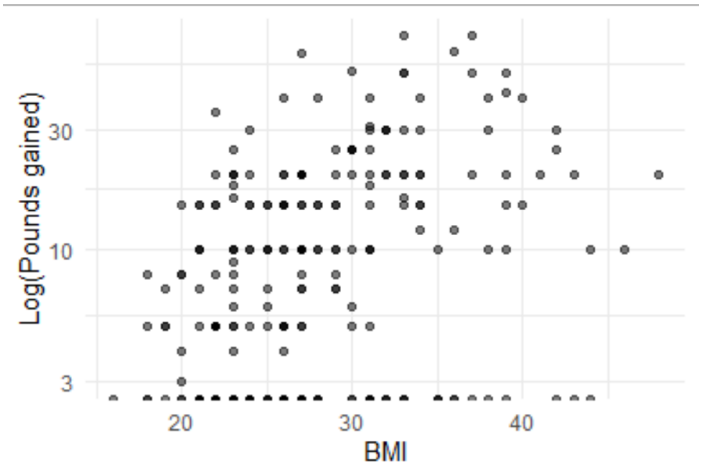


Figure 6: BMI vs pounds gained



In Figure 6, we see a positive correlation between BMI and pounds gained, after applying a logarithmic scale to pounds gained due to right-skewness. This indicates that as the starting BMI increases, pounds gained also increases.

MISSING DATA

Missing data was a substantial issue for the data used in this analysis, and there were two types of missingness present in the data: overall survey nonresponse and item-level missingness. Missing data only threatens the accuracy of statistical results if the missingness is non-random - meaning that there are one or more unobserved, external factor(s) that influence whether a respondent completes the survey. There are several analytic steps that can be taken to address missing data, but, first, it is important to examine patterns of missing data, which can include examining which questions are more likely to have missing values and the number of missing values within individual cases.

While this survey was sent to approximately 1,100 employees working at the call center, only 352 respondents returned surveys that were at least partially complete, which is a response rate of about 32%. Unfortunately, we do not have data on the demographic composition of the call center as a whole so we cannot examine whether there were systematic differences in which employees were more likely to answer the survey. Consequently, for the purposes of this study, we have to assume that the 352 respondents in our sample are a random sample that is representative of all employees working at the call center.

Two variables, metabolic exercise time and number of pounds gained were missing on items that could be calculated from other variables in the study. In 97 cases, the total metabolic exercise time was missing, while the component parts (vigorous exercise time, moderate exercise time and walking exercise time) were present in the data. Using the formula for the total metabolic exercise time given to us, we filled in the missing data for those cases. Furthermore, the number of pounds gained was only asked of respondents who responded “yes” to the question of whether or not they gained weight over the study period. For the pounds gained question, we make the assumption that respondents did not lose weight, and replaced the missing values for respondents who reported not gaining weight with 0 pounds ($N = 110$). While we might lose some specificity

by assuming that respondents did not lose weight, this assumption will not inhibit our analysis, because we are interested in weight *gain* rather than weight *loss*.

Table 2 displays the item-level missingness for each variable considered in our analysis. There were 114 cases (32% of the sample) with missing data on at least one analysis variable. Because we are considering two specifications of the dependent variable with different numbers of missing cases, and individuals missing data on the dependent variable cannot be included in regression analyses even after techniques used to address missingness, the minimum analytic sample size will be different dependending on our choice of the dependent variable. There were 4 respondents missing information on whether they gained weight during the sample period and 10 respondents missing information on the number of pounds gained during the study period. The independent variable with the most missing data was the BMI variable, with 99 missing cases, and 30 respondents had missing information about their age.

Table 2: Item-level missing data

Variable	*Number of Missing Cases*
Age	30
Gender	5
BMI	99
Metabolic equivalent exercise level	1
Shift start time	4
Number of pounds gained (continuous specification)	10

Next, we examined missingness within individual cases (by respondent). Of the 114 respondents with missing data, 70 respondents were missing two items, while only eight respondents were missing between four to six items. This indicates that missingness was not a widespread problem within respondents and was a problem for only a very small proportion of the respondents.

We address the problem of missing data separately for our use of random forest techniques as well as for our regression analysis.

RANDOM FOREST MODEL

There are 3 general approaches that Random Forest uses to address missing data: Proximity imputation, on-the-fly imputation and Missforest algorithm[1]. We decided to choose the Missforest algorithm over the other two imputation algorithm for the following reasons:

1. Proximity imputation produces **biased** prediction error estimates[1, 2], meaning that the predictor importance scores, which are computed based on prediction error, will also be biased. This happens because Proximity imputation uses methods such as median imputation to pre-imputes missing data first, and then uses the imputed data to fit forest models. Since predictor importance scores measure the dependence between responses and predictors in random forest models, our interpretations that are based on the predictor

importance scores provided by proximity imputation will not be reliable. Missforest algorithm, on the other hand, doesn't use pre-impute data. It uses non-missing values to fit forests first, and then uses the forests to predict the values that are missing[1]. Essentially Missforest turns imputation problems into prediction problems. Since only non-missing values are used to do imputation, no biases are introduced in the imputation process.

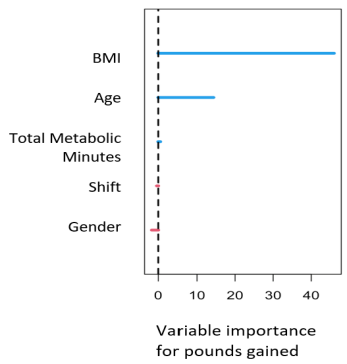
2. Proximity imputation can't deal with data that has missing values in responses[3]. Since pounds gained was our response and it contained 10 missing values, it would be awkward to implement proximity imputation in our data.
3. Even though both Missforest and on-the-fly methods produce unbiased prediction error estimates, Missforest generally gives us more accurate imputed data[2]. Because of that, the Random Forest will give us more accurate predictions when data is imputed by Missforest. The only downside of the Missforest method is that it is more computationally expensive than the other two methods, but since our data size is relatively small(about 2000 data points), the Missforest method can process our data fairly quickly. The time consumption will not be an issue in our analysis.

We will use the R package *randomForestSRC* by Hemant Ishwarant and Udaya Kogalur to perform random forest regression analysis. The function *impute.rfsrc* in this package will be used to perform the Missforest imputation algorithm on our data, and the function *rfsrc* will be used to fit Random Forest regression on the imputed data.

To keep the data consistent with the data we used to fit the regression models model, we also standardized the variable total metabolic exercise and categorized the variable shift into 4 categories, and the 4 categories in shift are *early.shift*(7am - 8am), *mid.morning*(9am - 11am), *early.afternoon*(12pm - 2pm), and *other*. In addition, We converted the gander to a factor variable(0 represents female and 1 represents male).

We started with inputting the data. We used *impute.rfsrc* from *RandomForestSRC* to apply the Missforest algorithm to our data. After that we used *rfsrc* to fit a Random Forest model on the data without missing values.

Figure 7: Variable importance in random forest model

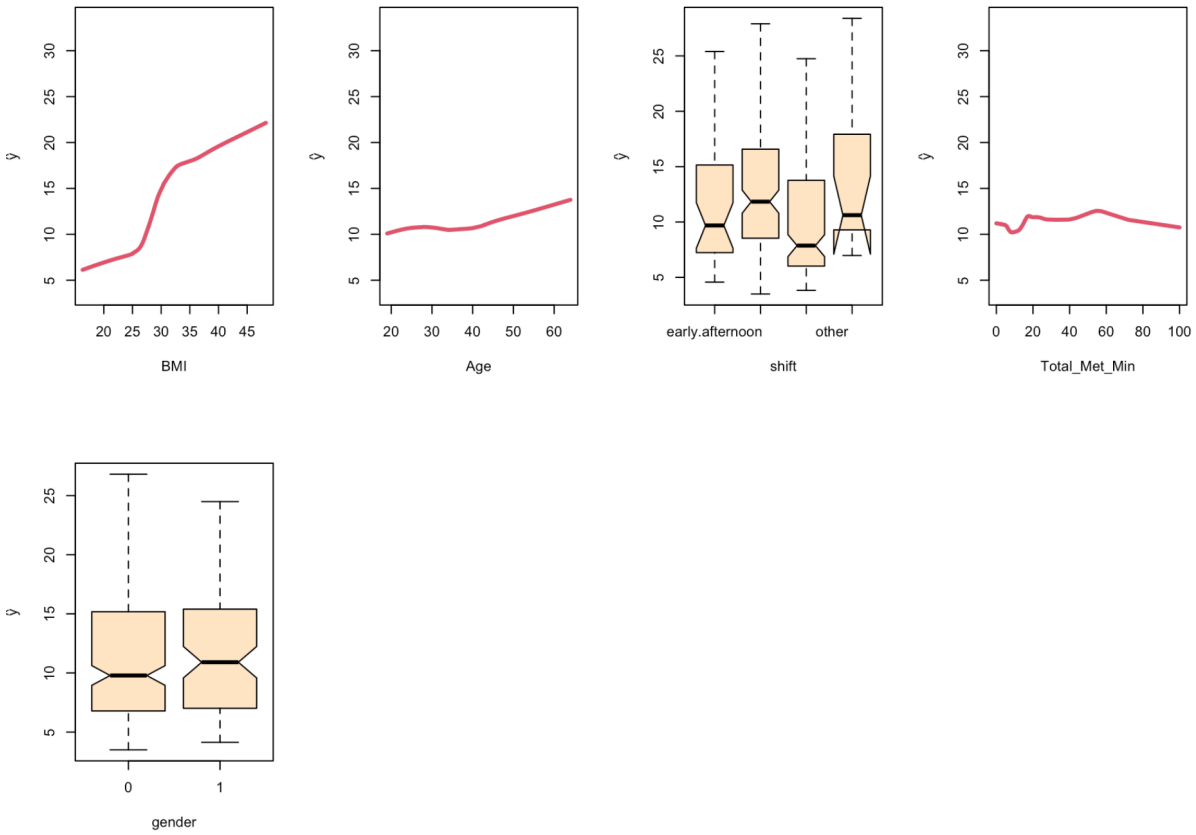


There are several ways to interpret the regression coefficients from random forest; one of them is the variance importance plot[4]. Variance importance plot shows the effects of predictors on response. The more important the score the predictor has, the stronger the effect of the predictor on response. The figure above is the variance importance plot from the random forest model. We can see that variable BMI is the most important predictors in this model,

and Age is the second most important predictor. This is consistent with the results from our regression analysis, discussed later. We also noticed that total metabolic minutes, shift and gender have low importance in this regression, indicating that they were not that significant. This is also consistent with what we found from later in our regression model.

Another way to illustrate the relationship between responses and predictors is Marginal effect plot[4]. Marginal effect plots fix all but one specific predictor, then vary the chosen predictor and observe how the prediction value reacts. The plots above show marginal effects of every predictor variable in our model. As we can see, prediction values vary the most when we vary the predictor BMI and Age. More specifically, both predictors have positive correlation with pound gain. This means that people with old age or high body mass index tend to gain more weight. total metabolic minutes, gender and shift barely affect the prediction value. This shows that BMI and age were the most significant factors in this model.

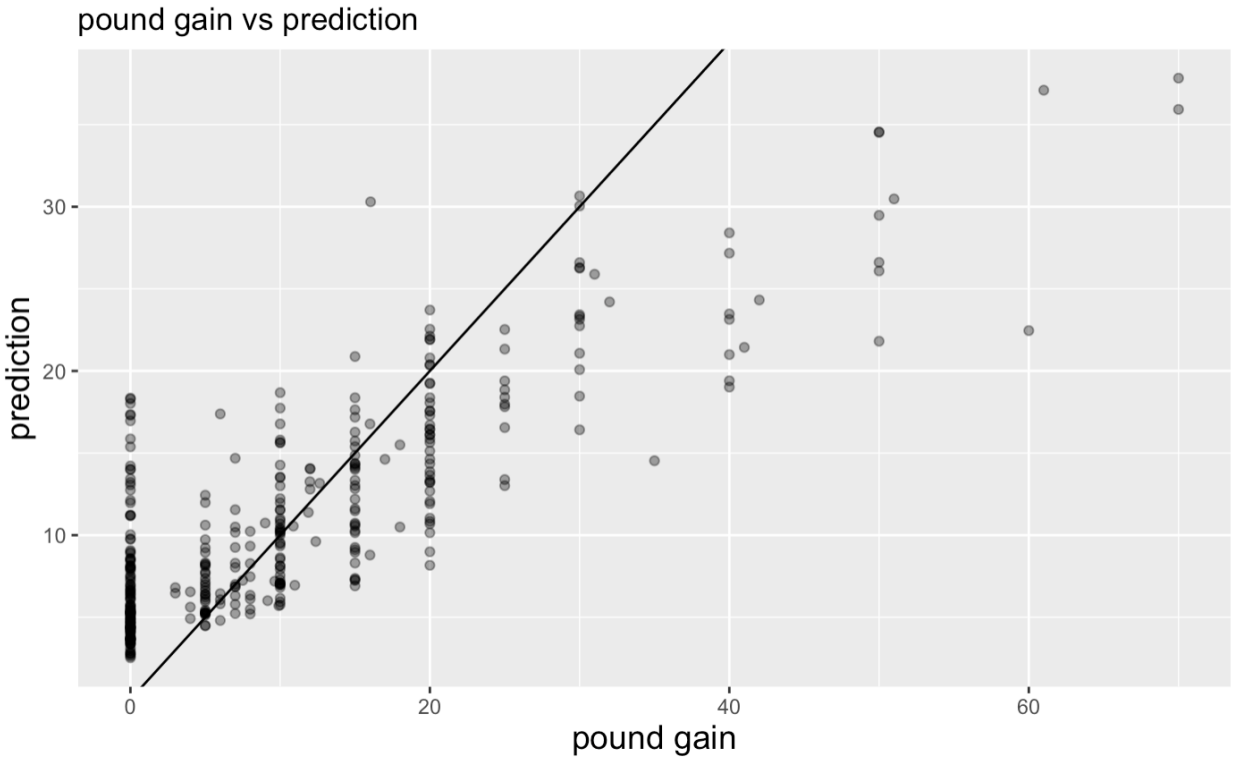
Figure 8: Marginal effect plots, Y hat represents the prediction value



Lastly, we will examine how well our model fits the data. Figure 9 shows the comparison between predicted values and pounds gained. In general, there is positive a correlation between prediction values and the observed values, however, our model tends to overestimate the amount

of weight gain when subjects gained little weight, and underestimate the amount of weight gain when subjects gained substantial amounts of weight. This shows that our model tends to give inaccurate predictions in extreme cases.

Figure 9: Pounds gained vs predicted values for random forest with 45-degree line.



FULL INFORMATION MAXIMUM LIKELIHOOD MODELS

Full information maximum likelihood (FIML) estimation is a straightforward method for addressing missing data that is easily implemented for linear regression. FIML produces efficient and unbiased estimates by using maximum likelihood estimation by using all available data in the dataset. Rather than dropping cases with missing data, it allows the estimates produced with incomplete cases to "steer" or push the parameters estimated using the cases with complete cases. Furthermore, unlike other methods of handling missing data, such as multiple imputation, FIML does not introduce additional random variation and produces identical results every time. Consequently, FIML is an ideal option for our analysis of weight gain over the study period.

We made three changes to the variables before they were used in the regression models. First, we standardized total metabolic exercise, because the range of the variable is so wide (0-12,852). Second, we collapsed the categorical variable for shift into a four category variable. In this

variable, the first category represents respondents with early morning (7am-8am; *n*=146), mid-morning (9am-11am; *n*=150), early afternoon (12pm-2pm; *n*=37), or other start time (*n*=15). Sensitivity analyses including the variable specifying each individual start time by hour did not produce substantively different results (see Appendix A for regression results).

Third, we used Box-Cox transformations to examine potential transformations of the dependent variable (see Appendix B for the statistical output for these tests). Because the dependent variable contains 0, we added 0.5 to the dependent variable before running the Box-Cox tests as the dependent variable must be strictly positive for these tests. The restricted log-likelihood test for the un-transformed dependent variable indicated that a transformation of the dependent variable was necessary ($p<0.001$). The suggested transformation of the dependent variable was 0.24, which suggests either a log transformation or a square root transformation. Because there are a substantial number of 0s in the dependent variable. We elect to use a square root transformation.

Following these adjustments to the data, we are now ready to estimate the relationship between BMI, exercise, and shift time on weight gain. To do this, we used a full information maximum likelihood linear model. We included quadratic specifications of age and BMI to allow for a curvilinear relationship between these variables and pounds gained. Although the magnitude the age^2 rounds to 0 in the final model, age is significant only when age^2 is included in the model, indicating that there is a slight curvilinear relationship between age and pounds gained, improving our understanding of the relationship between age and pounds gained. The estimate for the quadratic of BMI also rounds to 0, but, in models that exclude the BMI^2, BMI has no substantively meaningful effect on pounds gained, which similarly indicates the usefulness of including the quadratic term.

Table 3 displays the results of the final FIML model predicting the number of pounds gained during the study period. All estimates and standard errors have been squared to return to the original metric of measurement following the square root transformation of the pounds gained during model estimation. There was no significant difference in the number of pounds gained for male compared to female call center workers ($p=0.08$). Additionally, shift start time was not significantly associated with the number of pounds gained during the study period. Compared to workers who began their shift in the mid-morning hours (9am-11am), workers who started in the early morning, afternoon, or other time periods were not significantly more or less likely to gain weight during the study period. Furthermore, the respondent's total metabolic exercise time was not significantly related to weight gain ($p=0.47$). Although these factors, gender, shift time and metabolic exercise time, are not significant in this sample, there are several exogenous factors that could contribute to their lack of significance within this sample, which we will discuss in the limitations section.

Table 3: FIML Regression Results Predicting Pounds Gained

	Regression Coefficient (Standard Error)
Male	0.18
	(0.06)
Shift start time (mid-morning shift reference)	
Early morning (7-8am)	0.18
	(0.06)
Early afternoon (12pm-2pm)	0.01
	(0.14)
Other shift time	0.39
	(0.30)
Age	0.04*
	(0.01)
Age^2	0.00*
	(0.00)
BMI	0.32**
	(0.03)
BMI^2	0.00**
	(0.00)
Metabolic exercise total	0.01
	(0.01)
Constant	10.87
	(6.94)
N	352

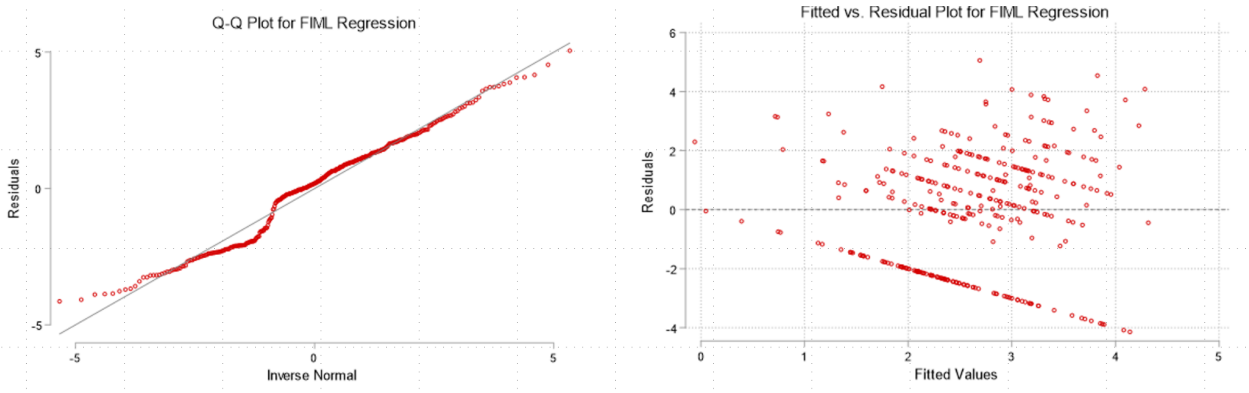
All regression coefficients and standard errors have been squared to adjust for the square root transformation of the dependent variable . * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Respondent age and BMI were the primary significant factors affecting weight gain during the study period. Older respondents were significantly more likely to gain weight during the study period (joint χ^2 test for the significance of age and age^2; χ^2=9.08, p=0.01). Each one year increase in age was associated with 0.04 greater pounds gained during the study period. Similarly respondents with higher BMIs were more likely to gain weight during the study period (joint χ^2 test for the significance of BMI and BMI^2; χ^2=39.50, p<0.001). Each one point increase in BMI was associated with 0.32 greater pounds gained during the study period.

Following analysis, we also created a series of graphs to examine the quality of model fit. Figures 1 and 2 display the QQ plot for normality and the fitted values vs. residual plot, respectively. The QQ plot looks relatively normal, with the exception of a dip around the middle of the graph. We suspect that is likely a result of the high proportion of respondents who gained no weight during the study period (32% of the sample). This fact of the data set is likely also responsible for the straight line toward the bottom left of the fitted values vs. residual plot. Although our data could theoretically exist along a continuous spectrum, the data used in this project was truncated at 0 and the number of pounds gained were largely rounded to the nearest pound. Consequently, any tests for normality in this dataset, will reflect the breaches of the

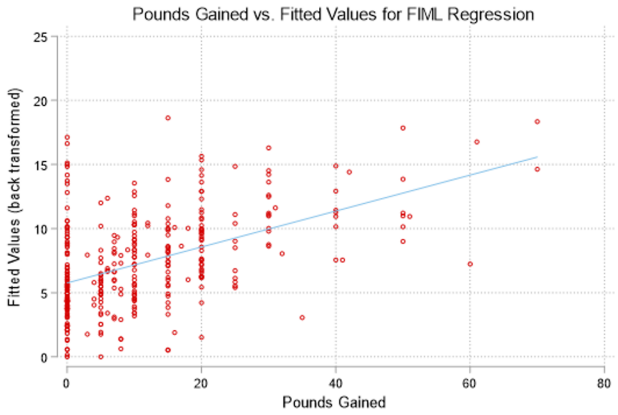
normality assumptions in our data. Nevertheless, the violations of normality do not seem to be dire, so we feel comfortable using a linear regression approach.

Figures 10 (left) and 11 (right): Q-Q and Residual Plot for FIML Regression



Lastly, we examined how well our final model predicts the number of pounds gained by comparing predicted pounds gained (fitted values) with the observed weight gain. Firstly, we see that our model underestimates the number of respondents who gained no weight during the study period. In general, there is a strong positive relationship between the observed values compared to the predicted pounds gained. Second, we see that the model underestimates the amount of weight gain for respondents who gained a substantial amount of weight in the survey period. However, this is to be expected as only around 13% of the sample gained more than 20 pounds during the survey period and models are less efficient around extreme cases.

Figure 12: Pounds gained vs fitted values for FIML regression



DISCUSSION

The results from both Random Forest model and FMIL show BMI and age are the most significant predictors in this study. More specifically, both factors are positively correlated with weight gain. This implies that older people and people with high body mass index are likely to

gain more weight. Predictors shift, total metabolic exercise time and gender show no significance in neither FMIL nor Random Forest.

Both models struggle to make accurate predictions for cases when subjects gained substantial amounts of weight. This is potentially due to the fact that the sample size of subjects who gained large amounts of weight is small. Also, both models tend to overestimate the amount of weight subjects gained when subjects gained 0 weight. This implies 0 is the hurdle and we may need to introduce a hurdle model such as a zero-inflated model to deal with the excess of 0 counts in the data or some thresholds in the model. For example, if the amount of weight gain prediction is less than 5 lbs, then the model treats the actual weight gain as 0 lbs.

Both models concluded that total metabolic exercise time, gender and shift don't affect the weight gain, however this conclusion may not be true. Our data was collected in a short time period, thus the factors that have long-term effects may not be significant in this study. It's very likely that factors such as shift time and metabolic exercise time have significant long-term impacts on weight gain. In the future, we need data that is collected over a long period of time to reveal the long-term effects of the factors that we are interested in.

Appendix A

Table 3: FIML Regression Results Predicting Pounds Gained with Shift Hour Specification

	Regression Coefficient (Standard Error)
Male	0.18
	(0.06)
Shift start time (7am shift reference)	
8am	0.13
	(0.18)
9am	0.05
	(0.22)
10am	0.02
	(0.24)
11am	0.10
	(0.25)
12pm	0.35
	(0.43)
1pm	0.68
	(0.66)
2pm	1.03
	(0.43)
Other shift time	0.25
	(0.43)
Age	0.03*
	(0.01)
Age^2	0.00
	(0.00)
BMI	0.33**
	(0.03)
BMI^2	0.00**
	(0.00)
Metabolic exercise total	0.01
	(0.01)
Constant	12.76
	(7.55)
N	352

All regression coefficients and standard errors have been squared to adjust for the square root transformation of the dependent variable . * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix B

```
 boxcox y gender shift3_em shift3_ea shift3_oth age age2 bmi bmi2 std_met, model(lhs)
Fitting comparison model

Iteration 0:   log likelihood = -957.23876
Iteration 1:   log likelihood = -848.96204
Iteration 2:   log likelihood = -825.81116
Iteration 3:   log likelihood = -825.35746
Iteration 4:   log likelihood = -825.35745

Fitting full model

Iteration 0:   log likelihood = -934.11738
Iteration 1:   log likelihood = -825.19748
Iteration 2:   log likelihood = -810.50625
Iteration 3:   log likelihood = -810.32406
Iteration 4:   log likelihood = -810.32406

                                                Number of obs   =        238
                                                LR chi2(9)      =      30.07
Log likelihood = -810.32406                     Prob > chi2     =      0.000

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      /theta |   .2361802   .0481705     4.90   0.000     .1417678    .3305925
------------------------------------------------------------------------------

Estimates of scale-variant parameters
------------------------------
             | Coefficient
-------------+--------------
Notrans      |
      gender |  -.5749394
   shift3_em |   .0188232
   shift3_ea |   .1876384
  shift3_oth |   .4257104
         age |  -.1505587
        age2 |   .0016938
         bmi |   .6070129
        bmi2 |  -.0081487
     std_met |  -.0864824
       _cons |  -4.775148
-------------+--------------
      /sigma |   2.163011
------------------------------

-------------------------------------------------------
   Test         Restricted     LR statistic
    H0:       log likelihood       chi2       Prob > chi2
-------------------------------------------------------
theta = -1      -1057.9579       495.27           0.000
theta =  0       -821.8896        23.13           0.000
theta =  1      -934.11738       247.59           0.000
-------------------------------------------------------
```

REFERENCE

1. Tang F, Ishwaran H. *Random Forest Missing Data Algorithm*. 2021 PubMed Central [link]
2. Ishwaran H, Kogalur U. *RandomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC).* R package version 2.12.1. 2021 CRAN [link]
3. Liaw A, Weiner M. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression.* R package version 4.6-14. 2021 CRAN [link]
4. Faraway J J. *Extending the Linear Model with R:Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Second Edition. 2016 Taylor & Francis Group. ISBN 978-1-4987-2098-4